

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2004-030093

(43)Date of publication of application : 29.01.2004

(51)Int.Cl.

G06F 17/30  
C12N 15/00  
// G01N 33/53  
G01N 37/00

(21)Application number : 2002-183810

(71)Applicant : HITACHI LTD

(22)Date of filing : 25.06.2002

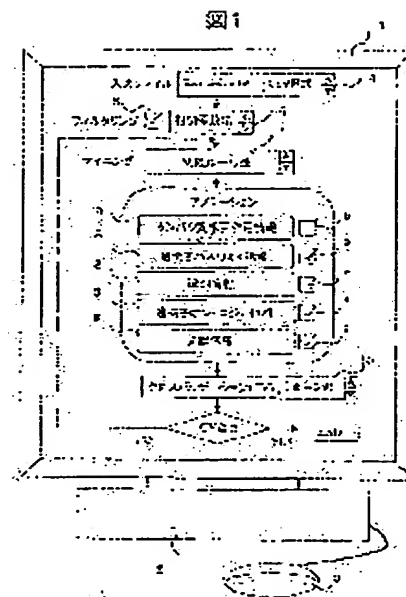
(72)Inventor : TOMITA HIROYUKI  
MAKI HIDEYUKI  
MORITA TOYOHISA  
SORIN SABAU  
TANIGAWA KOJI

## (54) METHOD FOR ANALYZING GENE EXPRESSION DATA

## (57)Abstract:

**PROBLEM TO BE SOLVED:** To provide a method and equipment for knowledge search based on gene expression data (also called a gene expression profile) using a DNA micro-array or the like.

**SOLUTION:** The knowledge search is done through: a process receiving the gene expression data; a process receiving class information; a process extracting a genetic group related to class classification by using a data mining technique; a process executing annotation with respect to the genetic group; a process extracting the common rule of the genetic group related to the class classification based on the genetic annotation; and a process executing the data mining using constraint conditions based on the common rule.



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

**This Page Blank (uspto)**

**Japanese Laid-Open Patent Publication**  
**No. 30093/2004 (*Tokukai* 2004-30093)**

A. Relevance of the Above-identified Document

The following is a partial English translation of exemplary portions of non-English language information that may be relevant to the issue of patentability of the claims of the present application.

B. Translation of the Relevant Passages of the Document

See also the attached English Abstract.

[INDUSTRIAL FIELD OF THE INVENTION]

The present invention relates to a method and an apparatus for finding information based on gene expression data (also referred to as “gene expression profile”) using a DNA microarray or the like.

[0002]

[PRIOR ART]

Research is underway to find gene functions through gene expression profile analysis and obtain information useful for drug development, pharmacology, toxicology, and diagnosis. For the analysis of DNA chip data, the following techniques have been used, for example: statistical analyses such as correlation analysis, principal component analysis, and analysis of variance; clustering such as k-mean clustering, hierarchical clustering, and self-organizing map; and classification algorithm such as nearest neighbor, discrimination analysis, support vector machine, neural network, and genetic algorithm (see Laura J. van't Veer et al., Gene expression profiling predicts clinical outcome of breast cancer, *Nature* 415, pp. 530-536 (2002); Scott L. Pomeroy et al., Prediction of central nervous

**This Page Blank (uspto)**

system embryonal tumor outcome based on gene expression, Nature 415, pp. 436-442 (2002)).

[0003]

[PROBLEMS TO BE SOLVED BY THE INVENTION]

However, there is no established means that can be used to simultaneously analyze large numbers of genes and obtain information. As to the information obtained by the analysis, it greatly depends on the knowledge of the analyzer. As such, the same information cannot always be obtained between different analyzers. Further, the amount of data obtained from the DNA chip is so large that it exceeds the analyzing capability of humans.

[0004]

An object of the present invention is to provide a method and an apparatus for analyzing a gene expression profile set. More specifically, the invention provides a technique for extracting, from a gene expression profile set, a group of genes that are useful for drug development, pharmacology, toxicology, and diagnosis, and a technique for finding laws common to such groups of useful genes.

[0005]

[MEANS TO SOLVE THE PROBLEMS]

In order to extract a group of useful genes from a gene expression profile set, and find laws common to such groups of useful genes, the invention performs the steps of (1) receiving gene expression data, (2) receiving class information, (3) extracting groups of genes associated with class classification, using a data mining method, (4) performing annotation on the groups of genes, (5) finding laws common to the groups of genes associated with class classification, based on gene annotation, and (6) performing data mining using restricting conditions that are based on the common laws. It is important that the

**This Page Blank (uspto)**

procedure from step (3) to step (6) be repeated. Conventional information systems employ steps (1) through (4). Conventionally, the groups of genes associated with class classification, and their annotations are presented to an analyzer in the form of, for example, a list or a graphical interface. The subsequent step of obtaining information relied on knowledge and intuition of the analyzer. A problem of this approach, however, is that analyzers have different knowledge levels. Another problem is that the amount of information obtained from the DNA chip gene expression profile is so large that the analysis takes a long time. Accordingly, there is a need for an information system for assisting an analyzer in obtaining information.

**This Page Blank (uspto)**



(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2004-30093

(P2004-30093A)

(43) 公開日 平成16年1月29日(2004.1.29)

(51) Int. Cl. <sup>7</sup>	F I	テーマコード (参考)
G 0 6 F 17/30	G 0 6 F 17/30 1 7 O F	5 B 0 7 5
C 1 2 N 15/00	G 0 6 F 17/30 2 2 O Z	
// G 0 1 N 33/53	C 1 2 N 15/00 Z	
G 0 1 N 37/00	G 0 1 N 33/53 M	
	G 0 1 N 37/00 1 0 2	
審査請求 未請求 請求項の数 17 O L (全 29 頁)		

(21) 出願番号	特願2002-183810 (P2002-183810)	(71) 出願人	000005108
(22) 出願日	平成14年6月25日 (2002. 6. 25)		株式会社日立製作所
			東京都千代田区神田駿河台四丁目6番地
		(74) 代理人	100075096
			弁理士 作田 康夫
		(72) 発明者	富田 裕之
			東京都千代田区神田駿河台四丁目6番地
			株式会社日立製作所ライフサイエンス推進
			事業部内
		(72) 発明者	牧 秀行
			神奈川県川崎市麻生区王禅寺1099番地
			株式会社日立製作所システム開発研究所
			内

最終頁に続く

(54) 【発明の名称】 遺伝子発現データ解析方法

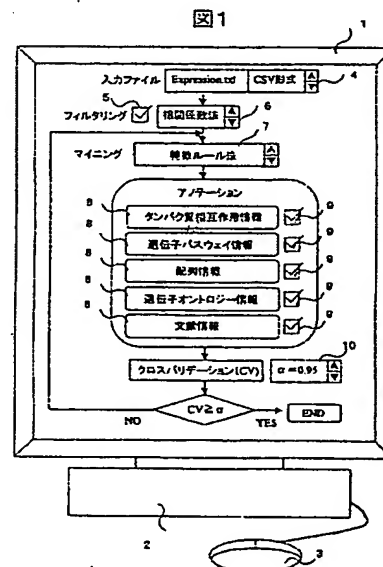
## (57) 【要約】

【課題】本発明の目的は、DNAマイクロアレイ等を用いた遺伝子発現データ（遺伝子発現プロファイルとも言う）に基づく、知識探索の方法および装置を提供することにある。

【解決手段】遺伝子発現データを受け取る工程、クラス情報を受け取る工程、データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程、前記遺伝子群にアノテーションを行う工程、遺伝子アノテーションに基づき前記クラス分類に関連する遺伝子群の共通規則を抽出する工程、前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程を行うことで知識探索を行う。

【効果】遺伝子発現データからメタ知識を得ることができる。

【選択図】 図1



**【特許請求の範囲】****【請求項 1】**

遺伝子発現データを受け取る工程と、クラス情報を受け取る工程と、データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程と、前記遺伝子群に注釈付け（アノテーション）を行う工程と、前記遺伝子アノテーションに基づき前記クラス分類に関連する遺伝子群の共通規則を抽出する工程と、前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程を有することを特徴とする遺伝子発現データ解析方法。

**【請求項 2】**

前記遺伝子発現データを受け取る工程の後、前記データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程の前に、遺伝子発現データを正規化する工程と、実験ばらつきを考慮して遺伝子発現データをフィルタリングする工程とを有することを特徴とする請求項 1 記載の遺伝子発現データ解析方法 10

**【請求項 3】**

前記クラス分類に関連する遺伝子群を抽出する工程において抽出された遺伝子群を用いてクロスバリデーションを行うことにより、該データマイニング手法を用いたクラス分類の正解率もしくはエラー率を比較する工程を、更に有することを特徴とする請求項 1 または 2 記載の遺伝子発現データ解析方法

**【請求項 4】**

前記データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程と、前記遺伝子群にアノテーションを行う工程と、前記遺伝子アノテーションに基づき前記クラス分類に関連する遺伝子群の共通規則を抽出する工程、前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程を、複数回繰り返すことを特徴とする請求項 1 乃至 3 何れかに記載の遺伝子発現データ解析方法。 20

**【請求項 5】**

前記クロスバリデーションにより得られる、該データマイニング手法を用いたクラス分類の正解率が、既定のしきい値を超えるまで、データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程と、前記遺伝子群にアノテーションを行う工程と、遺伝子アノテーションに基づき前記クラス分類に関連する遺伝子群の共通規則を抽出する工程と、前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程とを繰り返すことを特徴とする請求項 3 記載の遺伝子発現データ解析方法。 30

**【請求項 6】**

前記データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程とは、遺伝子データから、グループ間二乗和とグループ内二乗和の比率（BSS/WSS 比）もしくは特徴ルール法での正解率が、上位 10 個から 200 個の遺伝子を抽出する工程であることを特徴とする請求項 1 乃至 5 何れかに記載の遺伝子発現データ解析方法。

**【請求項 7】**

前記遺伝子アノテーションに基づきクラス分類に関連する遺伝子群の共通規則を抽出する工程は、前記クラス分類に関連する遺伝子群を抽出する工程により抽出された個々の遺伝子について、遺伝子間の相互関係である二項関係、パスウェイ、ゲノム、階層構造、ネットワーク関係のいずれか 1 つもしくは複数に属する共通ルールを検索する工程であることを特徴とする請求項 1 乃至 5 何れかに記載の遺伝子発現データ解析方法。 40

**【請求項 8】**

前記相互関係が二項関係のリガンドーレセプターの関係であり、前記抽出された遺伝子がリガンド遺伝子であれば、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件はレセプター遺伝子もしくはリガンド遺伝子とレセプター遺伝子の組を含む条件とし、前記抽出遺伝子がレセプター遺伝子であれば、前記拘束条件をリガンド遺伝子もしくはリガンド遺伝子とレセプター遺伝子の組を含む条件とすることを特徴とする請求項 7 記載の遺伝子発現データ解析方法。

**【請求項 9】**

前記相互関係がパスウェイであり、前記抽出遺伝子がパスウェイ PA 上にあり、クラス分 50

類に関連する遺伝子群の共通規則をパスウェイPA上の上流遺伝子、下流遺伝子、パスウェイPAと相関するパスウェイPB上の遺伝子、もしくは前記いずれかの遺伝子の組を含む条件とすることを特徴とする請求項7記載の遺伝子発現データ解析方法。

【請求項10】

前記相互関係がゲノムであり、前記抽出遺伝子が染色体CA上にあり、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件を染色体CA上の隣接遺伝子、もしくは前記抽出遺伝子と前記の隣接遺伝子との組を含む条件とすることを特徴とする請求項7記載の遺伝子発現データ解析方法。

【請求項11】

前記相互関係が階層構造でかつオントロジーであり、前記抽出遺伝子のオントロジーがOAであり、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件をオントロジーOAの上階層のオントロジーを有する遺伝子、もしくは前記抽出遺伝子と前記の上階層のオントロジーを有する遺伝子との組を含む条件とすることを特徴とする請求項7記載の遺伝子発現データ解析方法。

【請求項12】

前記相互関係が階層構造でかつ酵素(EC)であり、前記抽出遺伝子のEC番号(Enzyme Commission)がECAであれば、請求項1記載の、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件を酵素ECAの上階層に属する遺伝子、酵素ECAと同一グループに属する遺伝子もしくは前記いずれかの遺伝子の組を含む条件とすることを特徴とする請求項7記載の遺伝子発現データ解析方法。

【請求項13】

前記相互関係が階層構造でかつスーパーファミリーであり、前記抽出遺伝子のスーパーファミリーがSFAであり、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件を同一スーパーファミリーSFAに属する遺伝子、もしくは前記抽出遺伝子と前記の同一スーパーファミリーの属する遺伝子との組を含む条件とすることを特徴とする請求項7記載の遺伝子発現データ解析方法。

【請求項14】

前記相互関係がネットワークでかつ文献情報であり、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件を、文献情報により、前記抽出遺伝子との関連が予想される遺伝子、もしくは前記抽出遺伝子と前記の文献情報から関連が予想される遺伝子との組を含む条件とすることを特徴とする請求項7記載の遺伝子発現データ解析方法。

【請求項15】

前記相互関係がネットワークでかつ蛋白質相互作用である、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件を、蛋白質相互作用により、前記抽出遺伝子との関連が予想される遺伝子、もしくは前記抽出遺伝子と前記の蛋白質相互作用から関連が予想される遺伝子との組を含む条件とすることを特徴とする請求項7記載の遺伝子発現データ解析方法

【請求項16】

前記相互関係がネットワークでかつ代謝経路情報であり、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件を、代謝経路情報により、前記抽出遺伝子との関連が予想される遺伝子、もしくは前記抽出遺伝子と前記の代謝経路情報から関連が予想される遺伝子との組を含む条件とすることを特徴とする請求項7記載の遺伝子発現データ解析方法

【請求項17】

前記遺伝子アノテーションに基づきクラス分類に関連する遺伝子群の共通規則を抽出する工程は、クラス分類に関連する遺伝子群を抽出する工程により抽出された個々の遺伝子について、遺伝子全体像(ゲノム)、転写産物全体像(トランスクリプトーム)、タンパク質全体像(プロテオーム)、酵素全体像(エンザイモーム)、代謝全体像(メタボローム)、相互作用全体像(インタラクトーム)、時間的空間的局在全体像(ローカリゾーム)、表現型全体像(フェノーム)いずれか1つ以上の全体像内における相互作用内の共通ルールを検索する工程であることを特徴とする請求項1記載の遺伝子発現データ解析方法。

## 【発明の詳細な説明】

## 【0001】

## 【発明の属する技術分野】

本発明は、DNAマイクロアレイ等を用いた遺伝子発現データ（遺伝子発現プロファイルとも言う）に基づく、知識探索の方法および装置に関する。

## 【0002】

## 【従来の技術】

遺伝子発現プロファイルを解析することで遺伝子機能を明らかにし、創薬、薬理学、毒性学、診断に供する知見を得るための研究がなされている。例えば、相関分析 (Correlation Analysis)、主因子分析 (Principal Component Analysis)、分散分析 (Analysis of Variance) などの統計解析、k平均クラスタリング (k-mean Clustering)、階層クラスタリング (Hierarchical Clustering)、自己組織化マップ (Self-organizing Map) などのクラスタリング、最短近傍法 (Nearest Neighbor)、判別分析 (Discriminant Analysis)、サポートベクターマシン (Support Vector Machine)、ニューラルネットワーク (Neural Network)、遺伝的アルゴリズム (Genetic Algorithm) などの分類アルゴリズムをDNAチップデータの解析に適用した例がある。Laura J. van't Veerら、Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, pp. 530-536 (2002) や、Scott L. Pomeroyら、Prediction of central nervous system embryonal tumor outcome based on gene expression. Nature 415, pp. 436-442 (2002) を参照。

## 【0003】

## 【発明が解決しようとする課題】

しかしながら、多数の遺伝子の発現を同時に解析して知識を獲得する手段については、いまだに確立されたとはいえない状況である。特に知識獲得については、解析者の知識に依存することが大きいので、解析者によって得られる知識に差があること、またDNAチップから得られるデータ量が膨大であり、そもそも人間の解析能力を超えているという問題があった。

## 【0004】

本発明の目的は、遺伝子発現プロファイルのセットを解析するための方法および装置を提供することである。更に詳しくは、遺伝子発現プロファイルのセットから、創薬、薬理学、毒性学、診断に有用な遺伝子群の抽出およびそれら有用遺伝子群に共通する規則を抽出する技術を提供することを目的とする。

## 【0005】

## 【課題を解決するための手段】

遺伝子発現プロファイルのセットから有用な遺伝子群の抽出およびそれら有用遺伝子群に共通する規則を抽出するために、(1) 遺伝子発現データを受け取る工程、(2) クラス情報を受け取る工程、(3) データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程、(4) 前記遺伝子群にアノテーションを行う工程、(5) 遺伝子アノテーションに基づき前記クラス分類に関連する遺伝子群の共通規則を抽出する工程、(6) 前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程を行う。また前記(3)から(6)の工程を反復することが重要である。従来の情報システムは工程1から4までを行っていた。クラス分類に関連する遺伝子群とそのアノテーションについて、例えば一覧表、あるいはグラフィカルユーザーインターフェースを用いて、解析者に提示するに留めていた。その後の知識獲得の工程は、解析者の知識と直感に頼っていた。しかし、解析者ごとの知識レベルが異なること、そもそもDNAチップ遺伝子発現プロファイ

ルデータから得られる情報が膨大であるので、解析に多大の時間がかかることが問題となっている。そのため知識獲得においても、解析者を支援する情報システムが必要とされている。

#### 【0006】

遺伝子発現解析を除く分野でのデータマイニングは、一般にインスタンス数>>アトリビュート数の関係にある。例えばデパートの顧客データを例にすると、インスタンス数（顧客数）イコール数千から数万、アトリビュート数（性別、年齢、年収など）イコールたかだか百個であり、上記の関係が成り立つ。しかしDNAチップでは反対に、インスタンス数（サンプル数：数個から数十）に対し、アトリビュート数（遺伝子数：数千から数万）が桁違いに大きいことが特徴である。この種の問題では、「高度なマイニング手法を用いることなく、既存のサンプルに対してある程度正確な予測ができる」一方で「新規なサンプルに対しては予測に失敗することが多い」ことが知られている。この理由は、例えば10個の遺伝子で説明できなければ、100個、200個と遺伝子を増やすことでいつかは説明ができるためである。また実験誤差をあらかじめ除くことが難しいのでアトリビュート値の信頼性が低いこと、そして誤り値（Inaccurate values）が少しでも含まれているとマイニング結果が大きく変わってしまう（ロバストでない）ことも知られている。このため、適切なデータを選択すること（Data Selection）、および解析に悪影響を与える誤り値を除くこと（Data Cleansing）などの前処理が重要となる。現在、Data Selection、Data Cleansingの確立した方法はないが、後のKnowledge Discoveryの成否を10  
20  
30  
40  
50  
60  
70  
80  
90  
100  
110  
120  
130  
140  
150  
160  
170  
180  
190  
200  
210  
220  
230  
240  
250  
260  
270  
280  
290  
300  
310  
320  
330  
340  
350  
360  
370  
380  
390  
400  
410  
420  
430  
440  
450  
460  
470  
480  
490  
500  
510  
520  
530  
540  
550  
560  
570  
580  
590  
600  
610  
620  
630  
640  
650  
660  
670  
680  
690  
700  
710  
720  
730  
740  
750  
760  
770  
780  
790  
800  
810  
820  
830  
840  
850  
860  
870  
880  
890  
900  
910  
920  
930  
940  
950  
960  
970  
980  
990  
1000  
1010  
1020  
1030  
1040  
1050  
1060  
1070  
1080  
1090  
1100  
1110  
1120  
1130  
1140  
1150  
1160  
1170  
1180  
1190  
1200  
1210  
1220  
1230  
1240  
1250  
1260  
1270  
1280  
1290  
1300  
1310  
1320  
1330  
1340  
1350  
1360  
1370  
1380  
1390  
1400  
1410  
1420  
1430  
1440  
1450  
1460  
1470  
1480  
1490  
1500  
1510  
1520  
1530  
1540  
1550  
1560  
1570  
1580  
1590  
1600  
1610  
1620  
1630  
1640  
1650  
1660  
1670  
1680  
1690  
1700  
1710  
1720  
1730  
1740  
1750  
1760  
1770  
1780  
1790  
1800  
1810  
1820  
1830  
1840  
1850  
1860  
1870  
1880  
1890  
1900  
1910  
1920  
1930  
1940  
1950  
1960  
1970  
1980  
1990  
2000  
2010  
2020  
2030  
2040  
2050  
2060  
2070  
2080  
2090  
2100  
2110  
2120  
2130  
2140  
2150  
2160  
2170  
2180  
2190  
2200  
2210  
2220  
2230  
2240  
2250  
2260  
2270  
2280  
2290  
2300  
2310  
2320  
2330  
2340  
2350  
2360  
2370  
2380  
2390  
2400  
2410  
2420  
2430  
2440  
2450  
2460  
2470  
2480  
2490  
2500  
2510  
2520  
2530  
2540  
2550  
2560  
2570  
2580  
2590  
2600  
2610  
2620  
2630  
2640  
2650  
2660  
2670  
2680  
2690  
2700  
2710  
2720  
2730  
2740  
2750  
2760  
2770  
2780  
2790  
2800  
2810  
2820  
2830  
2840  
2850  
2860  
2870  
2880  
2890  
2900  
2910  
2920  
2930  
2940  
2950  
2960  
2970  
2980  
2990  
3000  
3010  
3020  
3030  
3040  
3050  
3060  
3070  
3080  
3090  
3100  
3110  
3120  
3130  
3140  
3150  
3160  
3170  
3180  
3190  
3200  
3210  
3220  
3230  
3240  
3250  
3260  
3270  
3280  
3290  
3300  
3310  
3320  
3330  
3340  
3350  
3360  
3370  
3380  
3390  
3400  
3410  
3420  
3430  
3440  
3450  
3460  
3470  
3480  
3490  
3500  
3510  
3520  
3530  
3540  
3550  
3560  
3570  
3580  
3590  
3600  
3610  
3620  
3630  
3640  
3650  
3660  
3670  
3680  
3690  
3700  
3710  
3720  
3730  
3740  
3750  
3760  
3770  
3780  
3790  
3800  
3810  
3820  
3830  
3840  
3850  
3860  
3870  
3880  
3890  
3900  
3910  
3920  
3930  
3940  
3950  
3960  
3970  
3980  
3990  
4000  
4010  
4020  
4030  
4040  
4050  
4060  
4070  
4080  
4090  
4100  
4110  
4120  
4130  
4140  
4150  
4160  
4170  
4180  
4190  
4200  
4210  
4220  
4230  
4240  
4250  
4260  
4270  
4280  
4290  
4300  
4310  
4320  
4330  
4340  
4350  
4360  
4370  
4380  
4390  
4400  
4410  
4420  
4430  
4440  
4450  
4460  
4470  
4480  
4490  
4500  
4510  
4520  
4530  
4540  
4550  
4560  
4570  
4580  
4590  
4600  
4610  
4620  
4630  
4640  
4650  
4660  
4670  
4680  
4690  
4700  
4710  
4720  
4730  
4740  
4750  
4760  
4770  
4780  
4790  
4800  
4810  
4820  
4830  
4840  
4850  
4860  
4870  
4880  
4890  
4900  
4910  
4920  
4930  
4940  
4950  
4960  
4970  
4980  
4990  
5000  
5010  
5020  
5030  
5040  
5050  
5060  
5070  
5080  
5090  
5100  
5110  
5120  
5130  
5140  
5150  
5160  
5170  
5180  
5190  
5200  
5210  
5220  
5230  
5240  
5250  
5260  
5270  
5280  
5290  
5300  
5310  
5320  
5330  
5340  
5350  
5360  
5370  
5380  
5390  
5400  
5410  
5420  
5430  
5440  
5450  
5460  
5470  
5480  
5490  
5500  
5510  
5520  
5530  
5540  
5550  
5560  
5570  
5580  
5590  
5600  
5610  
5620  
5630  
5640  
5650  
5660  
5670  
5680  
5690  
5700  
5710  
5720  
5730  
5740  
5750  
5760  
5770  
5780  
5790  
5800  
5810  
5820  
5830  
5840  
5850  
5860  
5870  
5880  
5890  
5900  
5910  
5920  
5930  
5940  
5950  
5960  
5970  
5980  
5990  
6000  
6010  
6020  
6030  
6040  
6050  
6060  
6070  
6080  
6090  
6100  
6110  
6120  
6130  
6140  
6150  
6160  
6170  
6180  
6190  
6200  
6210  
6220  
6230  
6240  
6250  
6260  
6270  
6280  
6290  
6300  
6310  
6320  
6330  
6340  
6350  
6360  
6370  
6380  
6390  
6400  
6410  
6420  
6430  
6440  
6450  
6460  
6470  
6480  
6490  
6500  
6510  
6520  
6530  
6540  
6550  
6560  
6570  
6580  
6590  
6600  
6610  
6620  
6630  
6640  
6650  
6660  
6670  
6680  
6690  
6700  
6710  
6720  
6730  
6740  
6750  
6760  
6770  
6780  
6790  
6800  
6810  
6820  
6830  
6840  
6850  
6860  
6870  
6880  
6890  
6900  
6910  
6920  
6930  
6940  
6950  
6960  
6970  
6980  
6990  
7000  
7010  
7020  
7030  
7040  
7050  
7060  
7070  
7080  
7090  
7100  
7110  
7120  
7130  
7140  
7150  
7160  
7170  
7180  
7190  
7200  
7210  
7220  
7230  
7240  
7250  
7260  
7270  
7280  
7290  
7300  
7310  
7320  
7330  
7340  
7350  
7360  
7370  
7380  
7390  
7400  
7410  
7420  
7430  
7440  
7450  
7460  
7470  
7480  
7490  
7500  
7510  
7520  
7530  
7540  
7550  
7560  
7570  
7580  
7590  
7600  
7610  
7620  
7630  
7640  
7650  
7660  
7670  
7680  
7690  
7700  
7710  
7720  
7730  
7740  
7750  
7760  
7770  
7780  
7790  
7800  
7810  
7820  
7830  
7840  
7850  
7860  
7870  
7880  
7890  
7900  
7910  
7920  
7930  
7940  
7950  
7960  
7970  
7980  
7990  
8000  
8010  
8020  
8030  
8040  
8050  
8060  
8070  
8080  
8090  
8100  
8110  
8120  
8130  
8140  
8150  
8160  
8170  
8180  
8190  
8200  
8210  
8220  
8230  
8240  
8250  
8260  
8270  
8280  
8290  
8300  
8310  
8320  
8330  
8340  
8350  
8360  
8370  
8380  
8390  
8400  
8410  
8420  
8430  
8440  
8450  
8460  
8470  
8480  
8490  
8500  
8510  
8520  
8530  
8540  
8550  
8560  
8570  
8580  
8590  
8600  
8610  
8620  
8630  
8640  
8650  
8660  
8670  
8680  
8690  
8700  
8710  
8720  
8730  
8740  
8750  
8760  
8770  
8780  
8790  
8800  
8810  
8820  
8830  
8840  
8850  
8860  
8870  
8880  
8890  
8900  
8910  
8920  
8930  
8940  
8950  
8960  
8970  
8980  
8990  
9000  
9010  
9020  
9030  
9040  
9050  
9060  
9070  
9080  
9090  
9100  
9110  
9120  
9130  
9140  
9150  
9160  
9170  
9180  
9190  
9200  
9210  
9220  
9230  
9240  
9250  
9260  
9270  
9280  
9290  
9300  
9310  
9320  
9330  
9340  
9350  
9360  
9370  
9380  
9390  
9400  
9410  
9420  
9430  
9440  
9450  
9460  
9470  
9480  
9490  
9500  
9510  
9520  
9530  
9540  
9550  
9560  
9570  
9580  
9590  
9600  
9610  
9620  
9630  
9640  
9650  
9660  
9670  
9680  
9690  
9700  
9710  
9720  
9730  
9740  
9750  
9760  
9770  
9780  
9790  
9800  
9810  
9820  
9830  
9840  
9850  
9860  
9870  
9880  
9890  
9900  
9910  
9920  
9930  
9940  
9950  
9960  
9970  
9980  
9990  
10000

#### 【0007】

遺伝子発現データとは、例えばDNAチップ法（DNA Chip）、ディファレンシャルディスプレイ法（Differential Display）、定量的PCR法（Quantitative PCR）、SAGE（Serial Analysis of Gene Expression）法、プロテインチップ法（Protein Chip）などの複数遺伝子もしくは蛋白質の発現変化を測定する方法により得られた複数遺伝子（あるいは蛋白質）に関する発現量、もしくは発現量同士の比率のことである。

#### 【0008】

クラス情報とは、例えばDNAチップ法等で測定された対象を分類するための情報である。例えば測定対象サンプルが被検査者の血液である場合、その血液が患者由来もしくは健康人由来のいずれであるかを、患者由来であれば1、健康人由来であれば0と定義する。また病理知見等に基づき、がんなどの疾患の悪性度を0, 1, 2, 3などと定義できる。培養細胞であれば、薬物投与前の細胞由来サンプルを0、薬物投与6時間後のサンプルを1、薬物投与12時間後のサンプルを2、薬物投与24時間後のサンプルを3などと定義できる。クラスが同一であれば、測定対象となる個人、個体が異なっても、同一の実験条件や性質、表現型（フェノタイプ）を有するものとする。

#### 【0009】

データマイニングとは、データベースから、興味ある規則性や因果関係を計算機で自動的に抽出する技術のことである。例えば決定木（Decision Tree）、ナイブベイズ（Naive Bayes）、フルベイジアン（Fully Bayesian）、相関ルール（Association rule）、特徴ルール（Characteristic rule）、EMクラスタリング（EM Clustering）、最短近傍法（Nearest Neighbor）、判別分析（Discriminant Analysis）、サポートベクターマシン（Support Vector Machi 50

ne)、遺伝的アルゴリズム (Genetic Algorithm)、線形回帰 (Linear Regression) のことである。もしくは上記方法を使用する上で、バギング (Bagging)、ブースティング (Boosting)、スタッキング (Stacking) などの手法を併用することもある。

#### 【0010】

アノテーションとは注釈付けのことである。例えば塩基配列、遺伝子機能情報、疾患関連情報、公共データベースの該当ID、他種遺伝子間 (例えばヒトとマウス) のホモログ情報、遺伝子ネットワーク情報、パスウェイ情報等である。

#### 【0011】

正規化 (ノーマライゼーションとも言う) とは、実験を行った時間、場所、作業者が異なることから実験のバックグラウンドや、ノイズの程度が異なるので、その実験ごとのバックグラウンドやノイズの程度を描える操作のことである。DNAチップであれば、チップごとに画像輝度 (蛍光強度) が異なることがあるので、まずバックグラウンド輝度、例えば、本来プローブが存在しない部位の輝度を取得して、そのバックグラウンド輝度の平均値や中央値を、プローブ蛍光強度から差し引くことなどの操作のことである。

#### 【0012】

フィルタリングとは、前記のように適切なデータを選択し、解析に悪影響を与える誤り値を除く操作のことである。例えばしきい値を定める方法がある。具体的には、信号強度が小さい (例えば500以下の) 場合、しきい値イコール500とし、500以下の信号強度を500にするかもしくはゼロにする操作のことである。

#### 【0013】

クロスバリデーション (Cross-Validation) とは、例えばテンフォールドクロスバリデーション (tenfold cross-validation)、もしくはリーブワンアウトクロスバリデーション (Leave-one-out cross-validation) のことである。テンフォールドクロスバリデーションとは、データセットをランダムに10個に等分割し、10分の9個のデータでトレーニングし、残りの10分の1個のデータでテストする合計10回の試行の結果から、正解率 (もしくはエラー率) を算出する方法である。リーブワンアウトクロスバリデーションとは、n個のデータセットのうち、n-1個のデータでトレーニングし、残りの1個のデータでテストする合計n回の試行の結果から、正解率 (もしくはエラー率) を算出する方法である。限られた数のデータセットから、各マイニング方法の正解率 (もしくはエラー率) を比較し、どのマイニング方法が優れているのかを定量的に評価する方法である。

#### 【0014】

図1は、本発明のソフトウェアを実行するために利用され得るコンピュータシステムと全体の工程を表したインターフェースの一例を示す。まず遺伝子発現データ、クラス情報が格納されているファイル入力のためにファイル形式やファイル名を入力する。続いて、正規化やフィルタリング方法の選択をし、データマイニングを行う。データマイニングにより抽出された遺伝子群にアノテーションをほどこす際に、どのアノテーションを行うかを選択する。またデータマイニングにより抽出された遺伝子群に基づいてクロスバリデーションを計算し、正解率がしきい値 ( $\alpha$ ; 図1では0.95) 以上ならば、そのまま終了し、しきい値以下であれば、アノテーション結果を反映させた拘束条件のもとで次のマイニングを行う。このマイニングは、正解率がしきい値を超えるまで自動的に反復して行える。

#### 【0015】

図2は、DNAチップの一般的な構造を示した図である。図16にDNAチップを用いた測定法のフローチャートを示す。まず支持体24にDNAプローブ22を固定化する。続いて、測定対象サンプルから抽出した遺伝子断片を蛍光標識などで標識する。この蛍光標識された遺伝子23を、DNAプローブ22とハイブリダイズさせる。その後、蛍光標識由来の蛍光を検出器21で検出する。この検出の結果、各DNAプローブ22にハイブリダイズした蛍光標識された遺伝子23の量が得られる。これを発現分布という。

## 【0016】

図3は、遺伝子アノテーションを行うに際して考慮すべき、遺伝子全体像（ゲノム：Genome）、転写産物全体像（トランスクリプトーム：Transcriptome）、蛋白質全体像（プロテオーム：Proteome）における遺伝子同士、転写産物同士、蛋白質同士、あるいは遺伝子と転写産物、転写産物と蛋白質、遺伝子と蛋白質の相互関係の例である。なお、ーム（ome）は全体あるいは全体像を意味する接尾語であり、遺伝子（Gene）の全体をゲノム（Genome）、転写産物（Transcript）の全体をトランスクリプトーム（Transcriptome）、蛋白質（Protein）の全体をプロテオーム（Proteome）と呼ぶ。本願明細書の以下の記述では、遺伝子、転写産物、蛋白質を遺伝子等と呼ぶ。図3における白丸が個々の遺伝子等を指し、白丸と白丸とをつないだ線は、実験等により既知となっている相互作用、因果関係である。図3（A）は遺伝子等に相互作用がない状態であり、独立と呼ばれる。但し将来、実験が行われることによってなんらかの相互作用、因果関係が発見される可能性はある。図3（B）は遺伝子等が1対1で相互作用する関係であり、二項関係と呼ばれる。例えば図4に示すように細胞表面に存在する受容体（レセプター）とその受容体に結合する結合体（リガンド）の関係が二項関係の一例である。図4（A）はインターロイキン2（IL2）蛋白質とインターロイキン2受容体アルファ（IL2RA）、インターロイキン2（IL2）蛋白質とインターロイキン2受容体ベータ（IL2RB）、インターロイキン2（IL2）蛋白質とインターロイキン2受容体ガンマ（IL2RG）、図4（B）は形質転換成長因子ベータ1（TGFB1）と形質転換成長因子ベータ受容体1（TGFB1R1）、形質転換成長因子ベータ1（TGFB1）と形質転換成長因子ベータ受容体2（TGFB1R2）、形質転換成長因子ベータ1（TGFB1）と形質転換成長因子ベータ受容体3（TGFB1R3）、図4（C）はエリスロポエチン（EPO）蛋白質とエリスロポエチン受容体（EPOR）の間のリガンドーレセプター関係である。またDNAとDNA結合蛋白質も二項関係の一例である。DNA結合蛋白質の例として転写因子、修復遺伝子などがある。図3（C）は道筋にそって遺伝子等同士が相互作用する関係でありパスウェイと呼ばれる。パスウェイ中には分岐は存在するものの、上流から下流へと一方向に相互作用が行われることが特徴である。パスウェイ上流とパスウェイ下流の遺伝子等の間には、因果関係が存在するともいえる。パスウェイの例として、図5に示すMAPキナーゼ（Mitogen Activated Protein Kinase）パスウェイのように、細胞表面の受容体とリガンドが結合したことを起点として、細胞表面から細胞内の細胞核にいたるまで情報を伝達するパスウェイが知られている。個々の丸印は遺伝子を、丸印と丸印をつなぐ矢印は、その矢印の方向に情報伝達が行われることを意味する。例えば図5のMos遺伝子からMEK遺伝子に、MEK遺伝子からERK遺伝子に情報が伝達される。パスウェイ情報としては、例えばパスウェイデータベース（<http://www.biocarta.com/>）を参照。

## 【0017】

図3（D）はDNA塩基配列上の遺伝子の相互配置関係であり、本願明細書ではゲノムと呼ぶ。ヒト、マウス、ラット等の高等動物では染色体上に、酵母や細菌などでは環状DNA上に遺伝子が配置されている。図6にゲノムの例を示す。ヒト第13番染色体上の13q12から13q13領域の一部では、LOC222428遺伝子からLOC160979遺伝子までが、図6のような位置に順番に存在している。ある疾患は染色体の一部の領域が欠失したり、増幅したりする結果、近傍にある遺伝子群が同時に欠失もしくは増幅することが原因で発症する。そのような遺伝子増幅、遺伝子欠失が原因となる疾患の原因遺伝子を探索するには、ゲノムの情報は有用である。図3（E）は遺伝子等同士が階層構造にある場合で、階層構造の例として図7のオントロジー、図8の酵素（EC：Enzyme Commission）、図9のスーパーファミリーの関係がある。図7はADPRTと呼ばれる遺伝子のオントロジーである。オントロジーとは遺伝子配列や蛋白質配列解析等に基づき、遺伝子の機能を定義した辞書である。オントロジーの詳細については、The gene ontology consortium、Gene ontology



: tool for the unification of biology. Nature Genetics 25, pp. 25-29 (2000) を参照。遺伝子オントロジーによるとADPRT遺伝子は、DNA修復 (DNA repair)、ADP-リボシル化 (ADP-ribosylation) の機能を有することが分かる。DNA修復はDNA代謝 (DNA metabolism) の一つであり、DNA代謝は核酸等代謝 (nucleobase, nucleoside, nucleotide and nucleic acid metabolism) の一つである。

#### 【0018】

図7のオントロジー右端括弧中の数字は登録遺伝子数を示す。386個の遺伝子がDNA修復機能を有する遺伝子として現在登録されており、DNA修復を含むDNA代謝には1138個の遺伝子が登録されている。遺伝子オントロジーでは汎用的な大分類から詳細な小分類へと階層構造を形成していることが、登録遺伝子数からも見て取れる。図8に階層構造の二つ目の例である酵素の例を示す。酵素はその構造や機能に基づいたEC番号 (Enzyme Commission) によって分類されている。EC1からEC6までであり、EC1はオキシドレダクターゼ (Oxidoreductases)、EC2はトランスフェラーゼ (Transferases)、EC3はハイドロラーゼ (Hydrolases)、EC4はリアーゼ (Lyases)、EC5はイソメラーゼ (Isomerases)、EC6はリガーゼ (Ligases) である。EC1から6は更に階層構造に従った分類がなされている。図8ではEC6の例をしめすが、EC6はEC6.1からEC6.5に分類される。EC6.3は更にEC6.3.1からEC6.3.5までに分類される。EC6.3.3の場合、実際の酵素はEC6.3.3.1からEC6.3.3.3である。

#### 【0019】

図9に階層構造の三つ目の例であるスーパーファミリーの例を示す。スーパーファミリーとは塩基配列の解析から得られるモチーフ、ドメイン構造から、類似の蛋白質立体構造や機能を有することが予想される一連の遺伝子群のことである。なおモチーフとは構造やパターンの要素のことであるが、ここでは各種のタンパク質のアミノ酸配列中に認められる一定の構造を指す。モチーフは互いに機能が異なる幅広いタンパク質に共通して見られる構造である。なおタンパク質にはドメインと呼ばれる構造があるが、タンパク質のドメインはモチーフの種々の組み合わせでできている。一般にモチーフは、タンパク質のドメインよりは小さい構造単位である。モチーフには、例えばヘリックス-スーターン-ヘリックスやジンクフィンガーと呼ばれるDNA結合構造モチーフなどがある。図9は薬物代謝酵素CYP遺伝子群の例を示している。CYP遺伝子群は全体で約50個ほど存在するが、それらの遺伝子群は、構造や機能ごとにCYP1A1, CYP1A2などのグループに分類されている。

#### 【0020】

図3 (F) は遺伝子等同士がネットワーク関係にある場合で、ネットワークの例として図10の文献情報、図11の蛋白質相互作用、図12の代謝経路の関係がある。図10にネットワークの一つ目の例である文献情報に基づく相互関係について示す。この方法は、二つの遺伝子名が同一の文献データベースの同一文章内に存在する数が多いほど両者の相互関係が強いとするものであり、例えばPubGeneデータベースを用いて相互関係スコアを得ることができる。Tor-Kristian Jenssenら、A literature network of human genes for high-throughput analysis of gene expression, Nature Genetics, vol. 28, pp21-28参照。図10の各丸印は遺伝子を、丸印同士をつなぐ線は、相互関係があることを示している。線に並んで記載された数字は相互関係スコアを表す。図10では相互関係スコアとは、線でつながれた二つの遺伝子が医学文献データベースMEDLINEの同一アブストラクト文中に存在した件数を示している。図10の中央にある遺伝子ADPRTと相互関係が強いものは、TP53, CFTR, EEF2, FRA1H, SP1, DAFの6遺伝子であり、相互関係ス



コアは6遺伝子とも1である。このPubGeneでは文献データベースとして米国NCBIのMEDLINEやOMIMを用いているが、その他の文献データベースでもかまわない。図11にネットワークの2つ目の例である蛋白質相互作用をしめす。米国カリフォルニア大学ロサンゼルス校のDIP (Database of Interacting Proteins) などのような、タンパク質相互作用データベースを用いて蛋白質相互作用を調べることができる。DIPについてはI. Xenariosら、DIP: the database of interacting proteins, Nucleic Acid Research, vol. 28, 289-291, 2000参照。タンパク質相互作用データベースにおいても、相互作用するタンパク質同士が線で結合されている。相互作用の強さは、例えば解離定数 (Dissociation Constant) が実験的に求められていれば、分子同士の結合力が分かるので、結合力の強い方をより相互作用が大きいと見なす。また1回の実験で確認された相互作用より、2回以上の複数回の実験で確認された相互作用を、より相互作用が強いと見なしでもよい。なおDIP以外のタンパク質相互作用データベースを用いてもかまわない。図11にネットワークの3つ目の例である代謝経路をしめす。代謝経路の詳細はKEGG (<http://www.kegg.kyoto-u.ac.jp>) などの代謝経路データベースを参照。図3(C)のパスウェイとこの代謝経路との違いは、補酵素 (co-enzyme) と呼ばれる反応を媒介する複数の蛋白質が代謝反応では関与するため、単純に上流から下流という一方向の関係にとどまらない複雑な構造をとる点である。

#### 【0021】

図3から図12はゲノム、トランスクリプトーム、プロテオームといった遺伝子情報における相互関係を示している。しかし遺伝子情報は生物情報の一部である。図13に示すように、生物情報には、酵素間相互作用 (酵素全体像: エンザイモーム: Enzymome, 代謝全体像: メタボローム: Metabolome)、相互作用 (相互作用全体像: インタラクトーム: Interactome)、時間的空間的局在 (局在全体像: ローカリゾーム: Localizome)、表現型 (表現型全体像: フェノーム: Phenome) にいたる階層構造がある。M. Vidal, A biological atlas of functional maps. Cell 104, pp. 333-339 (2001) を参照。本願明細書の解析法は、ゲノム、トランスクリプトーム、プロテオーム内の相互作用に留まらず、図13に示す生物情報全般における相互作用についても適用することができる。

#### 【0022】

図14と図15に本発明を実現するためのコンピュータ化された方法を示す模式図を示す。図14は、クロスバリデーション値と既定値とを比較することで繰り返し回数を決定する、本発明を実現するためのコンピュータ化された方法を示す模式図である。

#### 【0023】

図15は、全てのクロスバリデーション値を保存して比較する、本発明を実現するためのコンピュータ化された方法を示す模式図である。図14と図15において、遺伝子発現データを受け取る工程、クラス情報を受け取る工程、データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程、前記遺伝子群に注釈付け (アノテーション) を行う工程、遺伝子アノテーションに基づき前記クラス分類に関連する遺伝子群の共通規則を抽出する工程、前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程は共通である。図14と図15では遺伝子発現データを受け取った後に、正規化とフィルタリング処理を行ってもよいし行わなくてもよいが、特にフィルタリングはData Selection、Data Cleansingというデータマイニング結果を左右する工程なので行うことが望ましい。

#### 【0024】

DNAチップのデータ解析において有用なフィルタリングの例を表1に示す。フィルタリング法の1例であるしきい値法は、前述のように信号強度が小さい (例えば500以下の) 場合、しきい値イコール500とし、500以下の信号強度を500にするかもしくはは

ゼロにする操作のことである。しきい値の決定には経験に基づく方法、ノンパラメトリック検定、例えばWilcoxon符号付順位検定に基づく方法などがある。

【0025】

【表1】

## 表1

### フィルタリングの例

10

- |                               |
|-------------------------------|
| (1)しきい値法 (Threshold method)   |
| (2)相関係数法 (Correlation method) |
| (3)分散分析法 (Modified ANOVA)     |
| (4)ベイズ推定法 (Bayesian T-test)   |

20

フィルタリング法の1例である相関係数法は、まず(1)実データの全遺伝子に対し、複数サンプルから2個のサンプルを無作為に抽出して、相関係数を計算する。(2)続いて人工的に作成したランダムデータに対し、前記と同様に相関係数を計算する。相関係数にはピアソンの相関係数を用いる。(3)ランダムデータでの相関係数の確率分布と、実データでの相関係数との確率分布を比較する。(4)ランダムデータの分布から見て有意に大きい相関もしくは逆相関を有する遺伝子群のみをデータマイニングの入力として用いる方法である。

30

【0026】

フィルタリング法の1例である分散分析法は標準的なANOVA (Analysis of Variance) と似ている。但し標準的なANOVAは、データ同士が互いに独立であることを仮定して、F分布を判定に用いる。実際の遺伝子では、遺伝子データ同士が何らかの相関があることが十分予想されるので、通常のANOVAを用いることは誤りである。そこで、F分布に相当する分布を実データからブートストラップ法などを用いて模擬的に作成する方法が取られることが多い。

40

【0027】

フィルタリング法の1例であるベイズ推定法について説明する。以下Cy3とCy5という二色の蛍光色素で、2種類のサンプルを染色して同時にハイブリダイゼーションするDNAチップ実験を想定する。予め、同一RNAを二つに分けて、片方をCy3、もう片方をCy5で標識し、同一チップ上で競合ハイブリさせる実験(チップ枚数は5、10枚程度必要)を行うことを考える。この実験のデータは、Cy5/Cy3の平均値 $=c$ は1.0であるが、分散 $\sigma^2$ については未知である正規母集団 $N(c, \sigma^2)$ から大きさ $n$ の無作為標本 $\{Y_1, Y_2, \dots, Y_i, \dots, Y_n\}$ を抽出し、その観測値 $y = (y_1, y_2, \dots, y_i, \dots, y_n)$ を得たことに相当する。チップ間差、色素間差、ハントリンクの個人差等の誤差要因は、実験間で相関がないと考え

50

られるため  $\{Y_1, Y_2, \dots, Y_i, \dots, Y_n\}$  は相互に独立 (independently and identically distributed; i.i.d) と言える。

【0028】

【式1】

$$Y_i \sim \text{i.i.d. } N(c, \sigma^2) = N(1, \sigma^2) \quad (\text{式1})$$

10

本ベイズ推定の最終目的は、 $Y_i$  の  $\sigma^2$  を推定することにある。本推定により母集団  $N(1, \sigma^2)$  の  $\sigma^2$  を推定すれば、式1より、 $Y_i$  の  $\sigma^2$  も同一である。ベイズ推定では  $\sigma^2$  は分布を持つとするので、あくまで  $\sigma^2$  の推定値が得られる。推定値としては、平均、モード（度数が最大となる点）や、信頼区間として最高密度信頼区間 (Highest Density Region; HDR) が得られる。90%最高密度区間は、いわば未知母数  $\sigma^2$  の90%区間のうちで最も短く、また事後分布のピーク値（事後モード）を必ず含み、かつ区間の両端における事後密度が等しくなるものである。

式1が成り立つとき、 $\{Y_1, Y_2, \dots, Y_i, \dots, Y_n\}$  の同時確率密度分布 ( $a_1 < Y_1 \leq b_1, a_1 < Y_1 \leq b_1, \dots, a_n < Y_n \leq b_n$  が同時に満たされる確率密度分布)  $p(y' | c, \sigma^2)$  は、

【0029】

【式2】

$$p(y' | c, \sigma^2) = \prod_{i=1}^n (1 / \sqrt{2\pi\sigma^2}) \exp(-1/2\sigma^2 \times (y_i - c)^2) \quad (\text{式2})$$

のように正規分布を掛け合わせた形式で書ける。但し  $\Pi$  は積記号である。従って、観測値ベクトル  $y = (y_1, y_2, \dots, y_i, \dots, y_n)$  が与えられたときの尤度関数  $l(\sigma^2 | y)$  は、

【0030】

【式3】

$$l(\sigma^2 | y) \propto \prod_{i=1}^n (1 / \sigma) \exp(-1/2\sigma^2 \times (y_i - c)^2) \quad (\text{式3})$$

40

【式4】

$$l(\sigma^2 | y) \propto (\sigma^2)^{-n/2} \times \exp(-ns^2/2\sigma^2) \quad (\text{式4})$$

となる。ただし、 $s^2$  は母平均  $c$  を中心とした観測値分布であり、下記の式で表される。

【0031】

【式5】

50

$$s^2 = (1/n) \times \sum_{i=1}^n (y_i - c)^2 \quad (\text{式 5})$$

ここで、分散の事前分布  $p(\sigma^2)$  として、無情報事前分布 (noninformative prior distribution) を仮定する。この仮定は母数に関して無知であるとする事で、事前分布についての恣意性を出来る限り排除し、事後分布はできるだけデータによって支配されるようにする点で妥当である。無情報事前分布として、局所一様事前分布を用いるのが一般的である。局所一様分布とは、未知母数を二乗しようが、三乗しようが、対数値をとろうが、事前情報の漠然性を表すために少なくとも局所的には一様に分布するような分布のことである。具体的にはフィッシャー情報量の平方根に比例するように定めればよいことが分かっている。事前分布として局所一様分布を用いるならば、

【0032】

【式6】

$$p(\sigma^2) \propto \sigma^{-2} \quad (\text{式 6})$$

とすれば良い。つまり  $\sigma^2$  の事前分布  $p(\sigma^2)$  は、 $\sigma^{-2}$  すなわち定数とする。次に事後分布を考える。ベイズの定理より

【0033】

【式7】

$$p(\sigma^2 | y) \propto l(\sigma^2 | y) p(\sigma^2) \quad (\text{式 7})$$

が成立するので、

【0034】

【式8】

$$p(\sigma^2 | y) \propto (\sigma^2)^{-(n/2+1)} \times \exp(-ns^2/2\sigma^2) \quad (\text{式 8})$$

より、 $\sigma^2$  の事後分布  $p(\sigma^2 | y)$  は、 $\chi^{-2}(n, ns^2)$  と等しい分布になる。なお  $\chi^{-2}(\nu, \lambda)$  とは、尺度母数  $\lambda$  をもつ自由度  $\nu$  の逆カイ二乗分布と呼ばれる分布である。 $\chi^{-2}(\nu, \lambda)$  の平均は  $\lambda/(\nu-2)$ 、モード (度数が最大となる点) は  $\lambda/(\nu+2)$  となることが分かっているので、 $\sigma^2$  の点推定値として、

【0035】

【式9】

$$\text{平均値を基準とした場合: } \sigma^2 = ns^2 / (n-2) \quad (\text{式 9})$$

【式10】

モードを基準した場合： $\sigma^2 = ns^2 / (n + 2)$  (式10)

を考えることができる。また式8より事後的に、

【0036】

【式11】

$ns^2 / \sigma^2 \sim \chi^2(n)$  (式11)

10

の関係が得られる。式11の $\chi^2(n)$ とは、自由度 $n$ のカイ二乗分布である。ここでは、 $ns^2$ が固定値(観測値)、 $\sigma^2$ が確率変数になっている。式11と数表を用いることで、HDRを求めることができる。式9, 10, 11をもちいることで、2種類のサンプル間の遺伝子発現比率データが得られたとき、その比率が1.0と比較してどの程度、統計的に有意に異なるかを知ることができる。1.0より有意に異なる比率を有する遺伝子群のみをデータマイニングの入力として使用すれば良い。例えば5回の実験で、 $y_1 = 1.4$ 、 $y_2 = 0.89$ 、 $y_3 = 1.24$ 、 $y_4 = 0.91$ 、 $y_5 = 1.04$ が得られたとすると、 $s^2 = 0.04788$ である。平均値を基準とした点推定(式9)より、 $Y_i \sim N(1, 0.0798)$ となる。またモードを基準とした点推定より(式10)より、 $Y_i \sim N(1, 0.0342)$ となる。式11より、 $\sigma^2 \sim 0.2394 \chi^{-2}(5)$ 、となることから、数表を用いることで $\sigma^2$ の90%HDRは0.019-0.177となる。なお式1から式11までに示したベイズ推定法は統計手法の一つの方法であるので、ベイズ推定法以外の方法、例えば、ネイマン・ピアソン流の推定法を用いて同様の推定を行ってもよい。

20

【0037】

発現データを読み込んだ後、正規化、フィルタリングを行う、またクラス情報を読み取ること、データマイニングに輸入するためのデータ形式である発現マトリクスを作成することができる。発現マトリクス形式を図16に示す。遺伝子アノテーションを行(Raw)、サンプルアノテーションを列(Column)、対応する発現レベル(例えば前述のCy5とCy3の比率:  $Cy5 / Cy3$ )をマトリクスにした構造である。この発現マトリクス形式のサンプルアノテーション部位にクラス情報を付加することで、データマイニングに適した構造となる。この発現マトリクスは例えばCSV形式やタブ区切り形式などの形式でファイルに保存することができる。

30

【0038】

次のデータマイニング工程は、1回目であれば第1次データマイニング、以後マイニングを反復して行った(イテレーションした)場合その回数が $n$ 回目であれば $n$ 次データマイニング工程と呼ばれる。チップデータに適したマイニング方法を4つ表2に示した。いずれも教師付学習法(スーパーバイズドメソッド)において代表的な方法である。最短近傍法、判別分析についてはSplus等の統計計算パッケージ(Splusのclassパッケージ)を用いて実行することができる。またサポートベクターマシンについてもフリーソフトでSplusのクローンであるR言語のe1071パッケージを用いて実行することができる。次に特徴ルール法について説明する。

40

【0039】

【表2】

## 表2

## マイニングの例

## 教師付学習法 (Supervised Methods)

10

(1)最短近傍法 (Nearest Neighbor)

(2)判別分析 (Discriminant Analysis)

(3)サポートベクターマシン (Support vector machine)

(4)特徴ルール法 (Characteristic Rule)

20

特徴ルール法は特開平 8-77010 に、データ分析方法として開示されている。特徴ルール法では、複数の属性項目からなるサンプルの集合を対象データとする。全てのサンプルは互いに同一の属性項目を持つ。それぞれの属性項目が取り得る属性値はサンプル数と比較して少数の離散値であることが要求される。典型的には、3通り程度の記号値である。元の分析対象データが実数値データである場合、適当な境界で値の範囲を区切り、「大」「中」「小」といった記号値に置き換える等の方法で離散化する。

【0040】

特徴ルール法を実施する際には、属性項目の1個を選び、「結論項目」とする。また、結論項目の取り得る属性値のうちの1個を選び、「結論項目値」とする。また、その他の複数の属性項目を選び、「条件項目」とする。

【0041】

特徴ルール法では、「IF (条件部) THEN (結論部)」という形式の IF-THEN ルールを生成する。ルールの条件部は、条件項目とその属性値の組、すなわち述語であり、複数の述語が同時に条件部に現れることを許すが、典型的には3個程度に制限する。また、結論部は、結論項目と結論項目値からなる述語である。これにより、結論部の述語はただ1つに決定され、一方、条件部は様々な述語の組み合わせを取り得る。したがって、生成し得る IF-THEN ルールの数は、一般に多数になる。これら多数の IF-THEN ルールの中から、対象データの特徴をよく表している比較的少数のルールを探索することが特徴ルール法の目的である。各 IF-THEN ルールが対象データの特徴をどの程度よく表しているかを評価するために、以下の評価尺度を用いる。条件部を A (複数の述語の組み合わせを含む)、結論部を B とする、「IF A THEN B」というルールの評価尺度  $\mu(A \rightarrow B)$  を次式のように定義する。

$$\mu(A \rightarrow B) = P(A)^{-\beta} \times \log[P(B|A) / P(B)]$$

ここで、 $P(A)^{-\beta}$  は  $P(A)$  の  $\beta$  乗を意味する。 $P(A)$  は対象データの中で条件部 A が満足される確率、すなわち、対象データ全体の中で、A という条件を満たすサン

ブルの割合を表す。同様に、 $P(B)$  は結論部  $B$  が満足される確率、 $P(B|A)$  は、 $A$  を満たすという条件の下で結論部  $B$  が満足される条件付確率を表す。 $\beta$  は使用者が指定するパラメータで、0 以上、1 以下の実数値である。この評価尺度によって与えられる評価値が大きいルールほど、対象データの特徴をよく表していると思える。また、上記の評価尺度の定義式における  $P(A)$  をカバー率、 $P(B|A)$  をヒット率と呼び、これらは、取り出されたルールを使用者が解釈する際の手がかりとして用いられることがある。

#### 【0042】

生成され得る多数の IF-THEN ルールの中から、評価値の大きい、比較的少数のルールを取り出すアルゴリズムはいくつか考えられるが、「総当たり法」は、もっとも単純な方法の 1 つである。これは、取り出すルール数（例えば、10）をあらかじめ決めておき、そして、条件部に同時に現れる述語数の上限（例えば、3）を定め、その範囲内で可能な全ての IF-THEN ルールを生成、評価し、その中で評価値の大きい上位のルールを、あらかじめ定めた数だけ取り出すというものである。

#### 【0043】

特徴ルール法を、クラス分類に関する遺伝子群の抽出に用いるには、対象データにおいて、クラス分類に関する属性をルールの結論項目とし、遺伝子アノテーションに該当する属性を条件項目とする。これにより、クラス分類に関して重要な遺伝子を条件部に持つ IF-THEN ルールを得る。

#### 【0044】

データマイニングの結果、クラス分類において重要な（クラス分類を行う規則に関連する）遺伝子群を抽出するが、一般に正解率が高い順（エラー率の低い順）から、1 つ以上で 10 個から 200 個、好ましくは 20 個から 50 個の遺伝子を選択する。この重要遺伝子群抽出には前述の特徴ルール法が最も優れている。

#### 【0045】

特徴ルール法以外の例として、例えば BSS/WSS 比を用いる方法がある。なお BSS は Between-group sum of squares、WSS は Within-group sum of squares の略であり、遺伝子  $j$  の BSS/WSS 比は下記の式で定義される。

#### 【0046】

##### 【式 12】

$$BSS(j)/WSS(j) =$$

$$\frac{\sum_i \sum_k I(y_i = k)(m(x_{kj}) - m(x_{.j}))^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - m(x_{kj}))^2} \quad (\text{式 12})$$

但し、 $m(x_{.j})$  は全検体における遺伝子  $j$  の平均発現量、 $m(x_{kj})$  は、クラス  $k$  に属する検体における遺伝子  $j$  の平均発現量、 $x_{ij}$  は遺伝子発現データマトリクス、 $I(y_i = k)$  は、 $y_i = k$  のとき 1、それ以外は 0 となる関数である。式 12 で表される BSS/WSS 比が大きいほど、クラス内部の相違と比較してクラス間の相違が大きいので、クラシフィケーションの際には BSS/WSS 比が大きい遺伝子、例えば上位 20 から 50 個を用いることが望ましい。

#### 【0047】

次にデータマイニングにより抽出されたクラス分類において重要な遺伝子群についてアノテーション付けを行う。アノテーションには図 3 から図 13 までの様々な相互作用のうちいずれか 1 つ以上を行う。表 3 には例として遺伝子オントロジーを用いてアノテーションを行った例を示す。データマイニングの結果抽出された遺伝子群が Probe 1 から Probe 5 であった場合、もともと DNA チップ設計時に既知である遺伝子クラスター番号 (UniGene 番号)、塩基配列番号 (GenBank 番号)、遺伝子名に加え、遺伝

子アノテーション工程において、染色体番号や遺伝子オントロジーを公共データベース等から検索し、表3を作成することができる。

【0048】

【表3】

表3

## アノテーションの例

プローブ 番号	Unigene 番号	Genbank 番号	遺伝子名	染色体 番号	遺伝子オントロジー
Probe1	Mm.19904	AA413214	Bcl2-associated X protein	7	integral membrane protein apoptosis regulator apoptosis
Probe2	Mm.98	AA450909	proteasome (prosome, macropain) subunit, beta type 6	11	20S core proteasome peptidase, proteasome endopeptidase ubiquitin-dependent protein degradation
Probe3	Mm.5341	AA065510	defensin beta 1	8	extracellular antimicrobial peptide defense response, xenobiotic metabolism
Probe4	Mm.24816	AI643210	coagulation factor II (thrombin) receptor	13	membrane, integral membrane protein blood coagulation factor G-protein coupled receptor protein signaling pathway
Probe5	Mm.850	AA106360	signal recognition particle 14 kDa (homologous Alu RNA-binding protein)	2	signal recognition particle RNA binding protein targeting

続いて遺伝子アノテーションから共通性や規則性を抽出する工程を行う。表3の遺伝子群はあるクラス分類において重要であることがデータマイニングにより分かった遺伝子群であるとする。この表3の遺伝子群から共通した性質や特徴を抽出することが、遺伝子アノテーションから共通規則を抽出する工程の目的である。例えば、表3の遺伝子オントロジーに対して、複数の遺伝子において共通に見られるオントロジーがないかを検索する。するとProbe1とProbe4で、integral membrane proteinが共通して見られる。そこで表3からは、蛋白質として膜(integral membrane)に存在する遺伝子群がクラス分類に関わっているという規則が潜んでいる可能性を見出すことができる。次のデータマイニングでは、例えば、DNAチップ上に搭載されているプローブに対し、integral membrane protein (GO番号: 0016021) に対応するプローブという拘束条件を設けて、それに該当する遺伝子群のみでデータマイニング・クロスバリデーションすることで、integral membrane proteinがどの程度、クラス分類に寄与するかを定量的に把握することができる。またintegral membrane protein (G



O番号:0016021)の上階層は、membrane (GO番号:0016020)である。そこで別のデータマイニングでは、DNAチップ上に搭載されているプローブに対し、integral membrane protein (GO番号:0016021)もしくはmembrane (GO番号:0016020)に対応するプローブという拘束条件を設けて、それに該当する遺伝子群のみでデータマイニング・クロスバリデーションしてもよい。仮に“integral membrane proteinを含む”を拘束条件としてマイニングを行った場合と“membraneを含む”を拘束条件としてマイニングを行った場合とでクロスバリデーション結果(正解率もしくはエラー率)を比較した場合に、前者の正解率が後者より良ければクラス分類には、membraneでなく、integral membrane proteinが重要であることが分かる。このように遺伝子アノテーションから共通規則を抽出する工程とは、前記のオントロジーのように、図3から図13に示した相互関係を、表3のようなクラス分類の際に重要な遺伝子リストから検索する工程である。規則性抽出において、表4から表6に示したルールに従うことでより一般性の高い規則を見出すことができる。この見出された規則を拘束条件として、この拘束条件のもとにデータマイニングを行う工程が、次の、「前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程」である。この工程では、n次マイニングより得られた重要遺伝子群とそのアノテーションを用いて(n+1)次マイニングを行う。このようにマイニングを繰り返す意義は、ルール生成で得られた遺伝子群に、アノテーションによって情報を付加し、これらの情報を各遺伝子の属性と考え、共通する特徴をさらに解析する。更にこの解析サイクルを繰り返すことで共通する特徴を徹底して探索することができることである。なお、規則の発明に際し、リガンドーレセプタ関係についてはT. G. Graeber and D. Eisenberg, Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. Nature Genetics 29, pp. 295-300 (2001) を、蛋白質相互作用関係についてはH. Geら、Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. Nature Genetics 29, pp. 482-486 (2001)をそれぞれ参考にした。但し、前述の2件の公知例はどちらも、最初の1回目のマイニングに際し入力する発現マトリクスを改変する方法を提示しているのみである。本願明細書に記載されたように前回のマイニング結果を元に、更にマイニングを連続的に反復実行することはない。これでは一般性・頑強性(ロバストネス)に優れた知識獲得を行うことは困難である。本発明は、一般性・頑強性に優れた知識獲得を自動で行える点で公知技術とは根本的に異なっている。

【0049】

【表4】

表4

n次マイニングの結果を用いて(n+1)次マイニングを行う際のルール(その1)

相互関係の種類	一次マイニングで発見された関連候補遺伝子	二次マイニングにおける探索シード
二項関係	リガンド遺伝子	○レセプター遺伝子 ○リガンドレセプター対
	レセプター遺伝子	○リガンド遺伝子 ○リガンドレセプター対
パスウェイ	パスウェイPA上の遺伝子	○パスウェイPAの上流遺伝子 ○パスウェイPAの下流遺伝子 ○相関パスウェイPB上の遺伝子 ○上記いずれかと一次マイニング遺伝子とのペア
ゲノム	染色体CA上の遺伝子	○染色体CA上の隣接遺伝子 ○上記と一次マイニング遺伝子とのペア
階層構造	オントロジーがOAの遺伝子	○OAの上階層のオントロジーを有する遺伝子 ○上記と一次マイニング遺伝子とのペア

10

20

30

【表5】

## 表5

n次マイニングの結果を用いて(n+1)次マイニングを行う際のルール(その2)

相互関係の種類	一次マイニングで発見された関連候補遺伝子	二次マイニングにおける探索シード
階層構造	酵素ECAの遺伝子	<ul style="list-style-type: none"> <li>○酵素ECAの上階層の遺伝子</li> <li>○酵素ECAと同一グループに属する遺伝子</li> <li>○上記いずれかと一次マイニング遺伝子とのペア</li> </ul>
	スーパーファミリー-SFAに属する遺伝子	<ul style="list-style-type: none"> <li>○同スーパーファミリー-SFAに属する遺伝子(モチーフ、ドメイン構造が相互に類似)</li> <li>○上記と一次マイニング遺伝子とのペア</li> </ul>
ネットワーク	遺伝子A	<ul style="list-style-type: none"> <li>○文献情報から遺伝子Aとの関連が予想される遺伝子B</li> <li>○上記と一次マイニング遺伝子(遺伝子A)とのペア</li> </ul>

10

20

30

【表 6】

表6

n次マイニングの結果を用いて(n+1)次マイニングを行う際のルール(その3)

相互関係の種類	一次マイニングで発見された関連候補遺伝子	二次マイニングにおける探索シード
ネットワーク	遺伝子A	○蛋白質相互作用データから遺伝子Aとの関連が予想される遺伝子B ○上記と一次マイニング遺伝子(遺伝子A)とのペア
	遺伝子A	○代謝経路情報から遺伝子Aとの関連が予想される遺伝子B ○上記と一次マイニング遺伝子(遺伝子A)とのペア

10

20

表4から表6について説明する。表4から表6において遺伝子間の相互関係が二項関係でありかつリガンド-レセプターの関係であった場合、クラス分類関連抽出遺伝子がリガンド遺伝子であれば、アノテーションから抽出された共通性や規則性に基づく拘束条件をレセプター遺伝子もしくはリガンド遺伝子とレセプター遺伝子の組を含む条件とし、クラス分類関連抽出遺伝子がレセプター遺伝子であれば、前記拘束条件をリガンド遺伝子もしくは

30

【0050】

表4から表6において遺伝子間の相互関係がパスウェイであった場合、クラス分類関連抽出遺伝子がパスウェイPA上にあれば、アノテーションから抽出された共通性や規則性に基づく拘束条件をパスウェイPA上の上流遺伝子、下流遺伝子、パスウェイPAと関連するパスウェイPB上の遺伝子、もしくは前記いずれかの遺伝子の組を含む条件とする。

【0051】

表4から表6において遺伝子間の相互関係がゲノムであった場合、クラス分類関連抽出遺伝子が染色体CA上にあれば、アノテーションから抽出された共通性や規則性に基づく拘束条件を染色体CA上の隣接遺伝子、もしくは前記の抽出された遺伝子と前記の隣接遺伝子との組を含む条件とする。

40

【0052】

表4から表6において遺伝子間の相互関係が階層構造でかつオントロジーであった場合、クラス分類関連抽出遺伝子のオントロジーがOAであれば、アノテーションから抽出された共通性や規則性に基づく拘束条件をオントロジーOAの上階層のオントロジーを有する遺伝子、もしくは前記の抽出された遺伝子と前記の上階層のオントロジーを有する遺伝子との組を含む条件とする。

【0053】

表4から表6において遺伝子間の相互関係が階層構造でかつ酵素(EC)であった場合、クラス分類関連抽出遺伝子のEC番号がECAであれば、アノテーションから抽出された

50

共通性や規則性に基づく拘束条件を酵素E C Aの上階層に属する遺伝子、酵素E C Aと同一グループに属する遺伝子もしくは前記いずれかの遺伝子の組を含む条件とする。

#### 【0054】

表4から表6において遺伝子間の相互関係が階層構造でかつスーパーファミリーであった場合、クラス分類関連抽出遺伝子のスーパーファミリーがS F Aであれば、アノテーションから抽出された共通性や規則性に基づく拘束条件を同一スーパーファミリーS F Aに属する遺伝子、もしくは前記の抽出された遺伝子と前記の同一スーパーファミリーの属する遺伝子との組を含む条件とする。

#### 【0055】

表4から表6において遺伝子間の相互関係がネットワークでかつ文献情報であった場合、アノテーションから抽出された共通性や規則性に基づく拘束条件とは、文献情報により、クラス分類関連抽出遺伝子との関連が予想される遺伝子、もしくはクラス分類関連抽出遺伝子と前記の文献情報から関連が予想される遺伝子との組を含む条件とする。

#### 【0056】

表4から表6において遺伝子間の相互関係がネットワークでかつ蛋白質相互作用であった場合、アノテーションから抽出された共通性や規則性に基づく拘束条件とは、蛋白質相互作用により、クラス分類関連抽出遺伝子との関連が予想される遺伝子、もしくはクラス分類関連抽出遺伝子と前記の蛋白質相互作用から関連が予想される遺伝子との組を含む条件とする。

#### 【0057】

表4から表6において遺伝子間の相互関係がネットワークでかつ代謝経路情報であった場合、アノテーションから抽出された共通性や規則性に基づく拘束条件とは、代謝経路情報により、クラス分類関連抽出遺伝子との関連が予想される遺伝子、もしくはクラス分類関連抽出遺伝子と前記の代謝経路情報から関連が予想される遺伝子との組を含む条件とする。

。本発明において、図14、図15、表4から6に開示した方法を用いて反復マイニングを行うことで、より一般性や頑強性（ロバストネス）が高いメタ知識（meta-knowledge）を自動で得ることができる。

#### 【0058】

##### 【発明の実施の形態】

公開データである急性白血病のデータセット（Golubら1999年）を使用し、具体的な発明の実施形態を示す。データファイルはMITのサイトからダウンロードした（<http://www.genome.wi.mit.edu/MPR>）。このデータセットは、2種類の急性白血病：ALL（acute lymphoblastic leukemia）、AML（acute myeloid leukemia）患者から血液を採取し、Affymetrixヒトチップ（6817遺伝子）を用いて発現分布を測定したものである。被験者数は、合計72人（内、38人がB-cell ALL、9人がT-cell ALL、25人がAML）である。このデータセット（72×6817）を用いて、発現データのみから2種類の白血病をクラシフィケーション（分類）することができるかを示す。以下、2種類の白血病（ALL、AML）をそれぞれ、クラス0、1と定義する。すなわち、ALL＝クラス0、AML＝クラス1とする。なお急性白血病のデータセット（Golubら1999年）は、72検体データを38のトレーニングセット（data\_set\_ALL\_AML\_train.txt）と、34のテストセット（data\_set\_ALL\_AML\_independent.txt）に分けて提供されている。そのため、ここでは38検体（トレーニングセット）を用いて学習をおこない、その結果を用いて34検体（テストセット）のクラシフィケーションを試みた。この分類の結果と、予め分かっている各検体のクラスとを比較することで正解率とエラー率を求めることができる。

#### 【0059】

以後特徴ルール法を用いてデータマイニングした例を示す。6818属性、72サンプル

のデータセットを対象データとした。このうちの6817属性は、DNAチップで測定した遺伝子1つ1つの発現量に対応し、残りの1属性は、白血病の種類（ALL、AML）に対応している。遺伝子発現量に対応する属性値は実数値データであるため、それぞれの属性項目ごとに「大」「中」「小」の3カテゴリに離散化した。この時、72個のサンプルが各カテゴリにほぼ均等に振り分けられるようにそれぞれの属性項目のカテゴリの境界を設定した。また、白血病の種類に対応する属性値はもともと2種類の離散値（クラス0、クラス1）なので、そのまま用いた。

#### 【0060】

遺伝子に対応する6817属性をIF-THENルールの条件項目、白血病の種類に対応する属性を結論項目、「クラス0」を結論項目値とし、また、条件部に現れる述語数の上限は1として、評価値の上位20個のルールを取り出した結果を表7に示す。表7の各行が1つのIF-THENルールに対応しており、評価値の大きい順に上から並んでいる。表7の第1列は、IF-THENルールの条件部に対応し、第2列、第3列、第4列はそれぞれ、ルールの評価値、ヒット率、カバー率に対応する。なお、ルールの結論部は全てのルールで「白血病の種類 = クラス0」で同一なので、表中では省略した。例えば、第1行は「IF U07139\_\_at = 大 THEN 白血病の種類 = クラス0」というルールであり、「U07139\_\_at」という遺伝子の発現量が「大」であるならば、白血病の種類はクラス0（ALL）となる傾向が大きいということを意味している。

#### 【0061】

ところで、第1、第2のルールの評価値はともに 0.148 で同じ、第3のルール以降は全て評価値 0.142 で同じである。このような評価値の同じルール同士の並び順には分析上の意味はなく、データファイルの中の属性項目の並び順に依存している。

#### 【0062】

#### 【表7】

表7 特徴ルール法を用いたデータマイニング例

条件部	ルール評価値	セット率	カバー率
U07139_at= 大	0.148	1.0	0.35
M94633_at= 大	0.148	1.0	0.35
U29176_at= 大	0.142	1.0	0.33
M28170_at= 大	0.142	1.0	0.33
U27460_at= 大	0.142	1.0	0.33
M31523_at= 大	0.142	1.0	0.33
X97267_rnal_s_at= 大	0.142	1.0	0.33
X85116_rnal_s_at= 小	0.142	1.0	0.33
M84371_rnal_s_at= 大	0.142	1.0	0.33
M12959_s_at= 大	0.142	1.0	0.33
L09209_s_at= 小	0.142	1.0	0.33
U22376_cds2_s_at= 大	0.142	1.0	0.33
D26156_s_at= 大	0.142	1.0	0.33
Z49194_at= 大	0.142	1.0	0.33
Y08612_at= 大	0.142	1.0	0.33
X99920_at= 大	0.142	1.0	0.33
X95735_at= 小	0.142	1.0	0.33
X93512_at= 大	0.142	1.0	0.33
X82240_rnal_at= 大	0.142	1.0	0.33
X68560_at= 大	0.142	1.0	0.33

10

20

## 【発明の効果】

30

本発明により、遺伝子発現データに基づく、知識探索を行うことができる。遺伝子等同士  
の相互作用アノテーションに基づく繰り返し探索を行うことで、より一般性や頑強性（ロ  
バストネス）が高いメタ知識（meta-knowledge）を自動で得ることができる。

## 【図面の簡単な説明】

【図1】本発明の実施形態のソフトウェアを実行するために利用されるコンピュータシス  
テムとその操作画面例。

【図2】DNAチップの概要図。

【図3】ゲノム、トランスクリプトーム、プロテオームにおける相互作用の例。

【図4】二項関係の一例であるリガンドーレセプタ関係の例。

40

【図5】パスウェイ関係の例。

【図6】ゲノム関係の例。

【図7】階層構造の一例である遺伝子オントロジー関係の例。

【図8】階層構造の一例である酵素（EC: Enzyme Commission）関係  
の例。

【図9】階層構造の一例であるスーパーファミリー関係の例。

【図10】ネットワークの一例である文献情報の例。

【図11】ネットワークの一例である遺伝子相互作用の例。

【図12】ネットワークの一例である代謝経路の例。

【図13】生物情報の階層構造の例。

50

【図14】クロスバリデーション値と既定値とを比較することで繰り返し回数を決定する、本発明を実現するためのコンピュータ化された方法を示す模式図である。

【図15】全てのクロスバリデーション値を保存して比較する、本発明を実現するためのコンピュータ化された方法を示す模式図である。

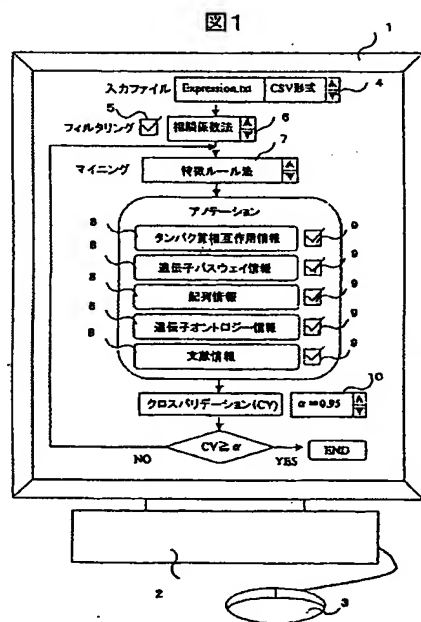
【図16】DNAチップを用いた測定法のフローチャートの一例である。

【図17】発現マトリクスの模式図である。

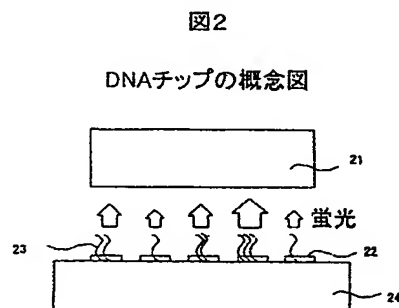
【符号の説明】

1. コンピュータシステム、2. CPU、3. 入力装置（マウス）、4. 入力ファイル名・形式選択、5. フィルタリング機能オンオフボタン、6. フィルタリングアルゴリズム選択、7. マイニング手法選択、8. アノテーション内容、9. アノテーション内容選択、10. イタレーション条件選択、21. 蛍光検出器、22. DNAプローブ、23. 蛍光標識された遺伝子、24. 支持体、41. リガンド、42. レセプター、51. 遺伝子、52. 遺伝子間バスウェイ、61. 染色体マップ、62. 遺伝子マップ、63. 遺伝子名、71. 遺伝子オントロジー（DNA repair）、72. 遺伝子オントロジー（protein amino acid ADP-ribosylation）、81. 酵素（EC）、91. 遺伝子スーパーファミリー、101. 遺伝子、102. 相互関係スコア、111. 蛋白質、112. 蛋白質間相互作用、121. 酵素、122. 代謝バスウェイ、123. 酵素反応生成物、171. サンプルアノテーション、172. 遺伝子アノテーション、173. 発現レベル。

【図1】



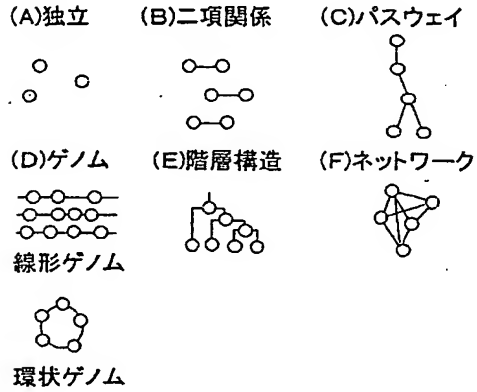
【図2】





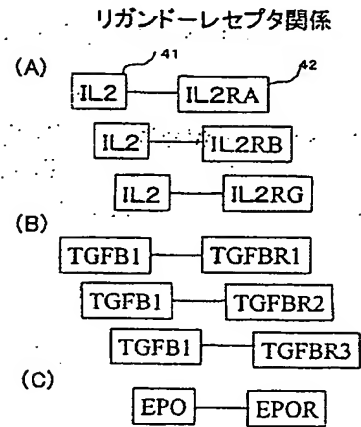
【図 3】

図3  
ゲノム、トランスクリプトーム、プロテオーム  
における相互関係の例



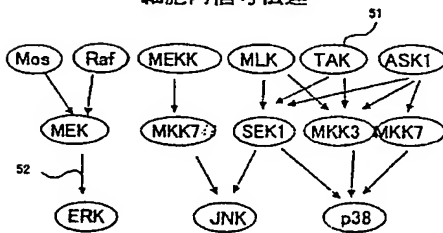
【図 4】

図4  
二項関係の例



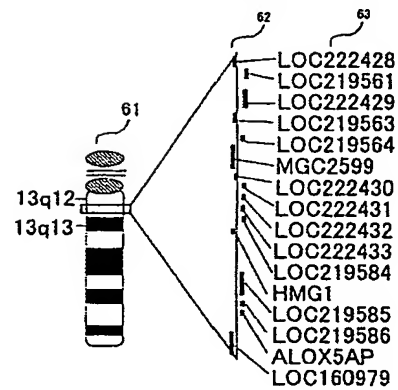
【図 5】

図5  
パスウェイの例  
細胞内信号伝達



【図 6】

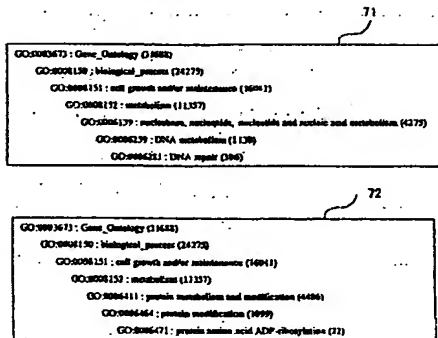
図6  
ゲノムの例



【図 7】

図7  
階層構造の例(その1)

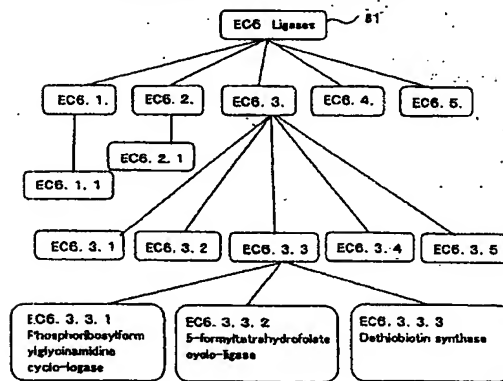
遺伝子オントロジー



【図 8】

図8  
階層構造の例(その2)

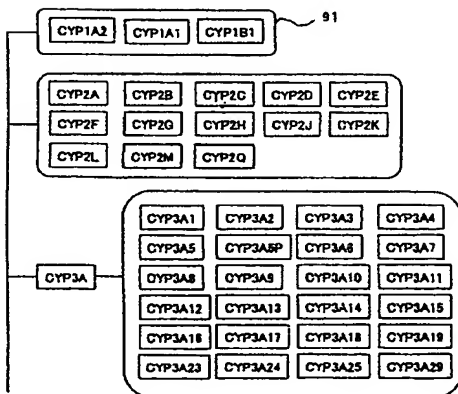
酵素(Enzyme Commission)



【図 9】

図9  
階層構造の例(その3)

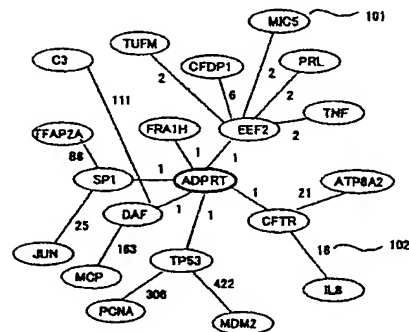
スーパーファミリー



【図 10】

図10  
ネットワークの例(その1)

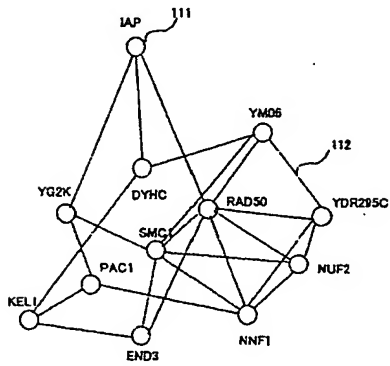
文献情報



【図 1 1】

図 11  
ネットワークの例(その2)

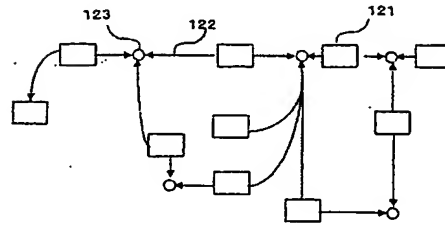
蛋白質相互作用



【図 1 2】

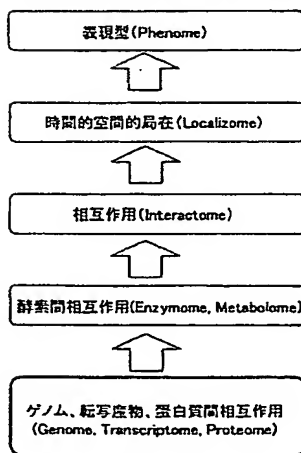
図 12  
ネットワークの例(その3)

代謝経路

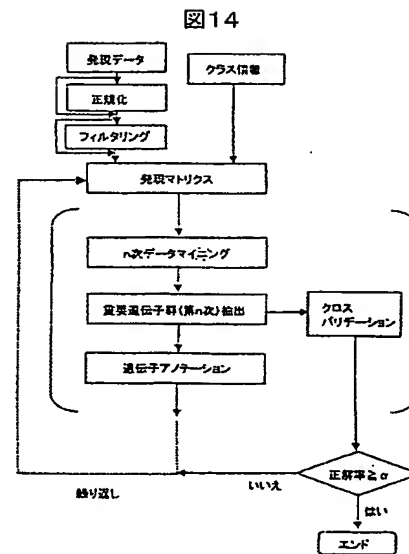


【図 1 3】

図 13  
生物情報の階層構造の例



【図 1 4】





---

フロントページの続き

(72)発明者 森田 豊久

神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所システム開発研究所内

(72)発明者 ソリン サバウ

東京都千代田区神田駿河台四丁目 6 番地 株式会社日立製作所ライフサイエンス推進事業部内

(72)発明者 谷川 浩司

東京都千代田区神田駿河台四丁目 6 番地 株式会社日立製作所ライフサイエンス推進事業部内

Fターム(参考) 5B075 ND20 NS10 UU19

**This Page Blank (uspto)**